·科学技术与社会·

人工智能道德增强的限度

The Limit of Artificial Intelligence Moral Enhancement

马翰林 /MA Hanlin

（苏州科技大学马克思主义学院，江苏苏州，215000）
(School of Marxism, Suzhou University of Science and Technology, Suzhou, Jiangsu, 215000)

摘　要：人工智能道德增强是从属于“人类道德增强”思潮的一个子话题——希望通过“人工道德建议者（AMA）”“客观”的道德表征能力与强大的信息获取能力协助人类提高道德认知水平。但由于AMA毕竟无法达到“理想观察者（IO）”的标准，其必然会陷入专家主义的窠臼。此外，诉诸构建“道德机器”的研究者们在通过学习算法构架自动道德分类与决策机器的时候也预设了某种IO视角。通过瓜里尼的道德案例分类器（MCC）的实验研究，我们发现无预分类的假设函数无法拟合，这意味着与IO契合的普遍主义AMA无法建成。鉴于此，我们建议另外一个不同的“道德增强”方案——一种“作为他者的AMA”，并寄希望于通用人工智能成为其实现基础。

关键词：人工智能　道德增强　理想观察者　通用人工智能

Abstract: Artificial intelligence moral enhancement is a sub-topic subordinate to the “human moral enhancement” trend. It is hoped that human moral cognition can be improved through the “artificial moral advisor (AMA)” , which is capable of “objective” moral representation and strong information acquisition. However, since AMA cannot meet the “ideal observer (IO)” standard after all, it will inevitably fall into the embarrassment of expertism. In addition, researchers who resort to the construction of “moral machines” also presuppose an IO perspective by constructing learning algorithmic frameworks for automatic ethical classification and decision making. Through the experimental study of Guarini’s Moral Case Classifier (MCC), we have found that the hypothetical function without pre-classification could not be fitted, which means that the generalist AMA that fits the IO cannot be built. In view of this, we propose another different “AI moral enhancement” scheme – an “AMA as The Other”, and hope that the artificial general intelligence (AGI) will become the basis for its realization.

Key Words: Artificial intelligence; Moral enhancement; Ideal observer; Artificial general intelligence

中图分类号：N0　　**文献标识码：**A　　DOI: 10.15994/j.1000-0763.2020.11.010

请设想一下如下情景：假设你是一名遵守社会公德的人，在没有做任何功课的情况下，生平第一次来到了另一个国家旅游。你发现这里的垃圾分类要求远远超过你的已有常识，因为你不知道你手中的某些垃圾的材质以及本地的分类要求是什么。这时你启动了手机里的智能对话系统，向它告知了自己的困境。它要求你把手中要处理的垃圾对手机摄像头，经过识别后，向你指定了最近的一个垃圾桶

基金项目：江苏省高校哲学社会科学研究基金项目“自由意志问题范式转换——计算主义自由意志研究”（项目编号：2018SJA1353）。
收稿时间：2019年3月27日
作者简介：马翰林（1985–）男，山东青岛人，苏州科技大学马克思主义学院讲师，研究方向为道德哲学、科技哲学。
Email: fanfan2011cn2000@aliyun.com

所在，并告知你应该把这个垃圾扔到左边数第二个桶中去。

这是一个在最近的未来很可能发生的“人工智能道德增强（AI Moral Enhancement）”的具体情景。该例子中人工智能的建议是被称作人工道德建议者（Artificial Moral Advisor, 简称AMA）[1]的道德AI（Moral AI）的功能之一。[2]AMA的目的不仅仅是为人类在具体情境中提供超出人类认知范围的道德相关信息，进而帮助人类做出更恰当的道德行为。它的终极目的甚至可以是帮助人类接近某种“理想观察者”的境界，即一个近乎全知的存在，从而做出一个建立在信息零缺失、（基于某类原则）“道德零失范”基础上的道德判断——这往往也是唯一正确的判断。作为一个新兴的、甚至暂时还有些冷门的话题，道德AI或AMA被展望和被讨论的程度还没有达到成熟的水平——尤其缺乏深入的质疑声，我们将尝试做这个工作，即试图证明这种AMA在“应该”和“能够”两个层面上都很难成立。

道德AI思路的出现象征着两个理论风潮的合流。一个风潮是作为第二波“人类增强”的讨论焦点的“人类道德增强”，[3]该理论的倡导者正是提出道德AI的朱利安·萨武列斯库（Julian Savulescu）等人，他们提议人类“应该”通过生理改造或其他外部调制（如人工智能辅助）提升人类的道德认知水平，更改人类产生“错误的”道德判断的生理或认知基础，从而应对道德危机所带来的全球性社会危机。[4]另外一个风潮指部分人工智能研究者希望制造某种“道德机器”。具体而言，其中一部分专用人工智能学者（如各类专家系统的开发者）致力于如何让机器“能够”合乎伦理地工作——如自动驾驶系统、自动手术系统或医疗反馈系统的伦理的控制论实现；而另外一部分人则被“创造类人的道德主体”这一话题所吸引，他们致力于创造一种具备各种类人特性的（如道德自主性）的通用型人工智能（AGI）——一种“思维机器”。[5]无论诉诸何种目的的道德机器研究，首先要解决的问题是“如何表征道德”。不同的技术路径对此问题的回答有所不同，这种差异在专用人工智能与部分通用人工智能学者之间表现得尤为显著。萨武列斯库明确认为通过“弱AI”技术——也就是专用人工智能系统就可以达成道德AI的目的。（[2]，p.84）然而这个目的可以达到吗？如果达不到却强行地推广道德AI又会导致什么样的伦理问题？要审视这些疑问，我们需要从如下几点逐步入手：

1. 理想观察者背后的深层道德哲学基础是什么？这是不是一个值得商榷的预设？

2. 理想观察者相关的道德哲学基础与什么样的道德机器表征体系是契合的？这样一个特定的道德机器技术本身存在哪些问题？

3. 这样的问题是否同理想化的道德AI可能造成的伦理问题相关？

一、AMA与理想观察者，绝对主义的道德指南？

为了避免误解，在讨论AMA之前有两个信息需要明确。首先，AMA不是一个简单的信息查询系统。虽然通过搜索引擎人们也可以查明如何在日本处理垃圾。但我们设想的是当使用者直接询问：“如何处理这个垃圾才最道德？”的时候，系统会识别该语句的道德含义并自动结合使用者的习惯来给出建议。这种系统在人类面临超出自己的信息检索能力的道德问题的时候往往会显示威力——如怎样为一个有待深度挖掘的社会道德事件定性？当某人想要做出一个独立而且比较完备的判断，而不希望被社会舆论左右的时候，他可以借助AMA强大的道德信息处理能力来下判断。其次，AMA只是道德AI中的一个类型，根据萨武列斯库等人的设想，（[2]，pp.85–89）道德AI至少包括四类道德增强功能：1. 道德环境监测器，提醒当事人自己的生理或者其他客观指标会对他的道德境况造成什么影响，如睡眠不足可能会导致人类的道德推理能力受损，[6]此时该系统将提醒当事人。2. 道德组织者，指为当事人既定的道德目标提供可行的方案。3. 道德提示器，当人类陷入道德抉择困难的时候，它会为当事人提供各类道德规范与标准，帮助其深化道德推理。4.道德建议者，即AMA。值得注意的是，一个完备的AMA往往同时也需要借助前三类功能的理论或技术基础。

之所以需要道德辅助增强，是因为普通人有如下缺陷：1.获取知识和信息的速度不够；2.推理速度不够；3.容易受到情感和情绪影响；4.不能像道德专家一样能够熟识各种道德规范。而一个理想的道德存在似乎应该没有这些缺陷，由罗德里克·弗斯（Roderick Firth）[7]在1952年提出的著名的理想观察者（ideal observer, 简称IO）被当做一

个备选来构建AMA，该模型的特征有:（1）对于非伦理事实无所不知（omniscient）;（2）无所不感（omnipercipient），能够通过视觉或想象等所有可能存在的方式认知这些信息;（3）不偏不倚;（4）不受情感左右;（5）始终如一;（6）其他如常。

之所以提出IO概念，是因为弗斯希望在经典经验主义（奎因所批评的那种）框架内提出一种不同于相对主义或情感主义道德哲学的语义分析理论——所谓的“绝对主义意向分析”（absolutist dispositional analysis, 简称ADA）。在ADA中，任何如“x是善的”的道德谓述都不是唯我论或心理主义的，而指称一类道德主语X，同时对接了针对x的某种可经验的道德意向。所以“x是善的”的意思是：x是这样一类道德存在，它的存在可以被给予某种肯定的道德意向。该谓述在逻辑上是“客观的”，因为它可以由一个IO的道德意向来表征，而IO的存在与否与该谓述的真假没有关系。换句话说，在ADA中一个命题的真假与否同它被意向的有意识的经验主体的存在与否没有关系，即命题的真假不依赖于相关意向主体的意识，可见这是一个类似于弗雷格体系的语义学系统，这种谓词逻辑系统通常是专用计算系统的基础预设。那么IO基于什么样的知识来判断一个道德语句的真假呢？弗斯采取了一种自然主义的解释，他认为任何一种道德属性都是一种“第二性质”,（[7]，p.324）关于某个语句表述的第二性质的判断的真假与否，必须通过更深的“第一性质”来判断，例如判断“水仙花是黄的”，我们只需要取下水仙花黄色的材质对其进行色素分析即可。同样，道德属性也可以被还原为客观的认知状态——例如行为状态或身体的状态来判定。所以在弗斯看来，作为IO的经验或道德判断证据的东西——所谓“道德数据”（[7]，p.326）不是道德意识的信念状态，因为后者是需要被判断的“第二性质”。由于IO具备完备的数据收集能力和无偏向的数据处理策略，所以它可以公正地对任何一个“x是善的”下判断。所以，IO被预设为一个完全确定的道德属性分类系统，而客观的状态被认为是这种分类的最好的表征。

在知识表征层面，AMA几乎全盘接受了ADA的设定。由于计算机在信息处理以及科学检测数据化方面的优势，它被设定为一个IO，用来对人类意识中的道德判断进行评估。AMA系统的目的是通过评估与建议，逐步协助人类接近IO的水平，或者慢慢被“训练”去接受来自接近IO处理能力的计算机的判断方式和倾向。由于IO的“道德数据”是“客观的”，所以道德心理学以及认知科学的相关科学性数据或其他超越人类检索能力的数据挖掘结果，都可以被用来作为AMA的判断依据，例如，“高水平的睾丸激素和睡眠不足会使人更容易做效应主义决策（会高估效用值）。AI会提醒行动者，由于这个原因，他对被建议的行动的评估可能与他惯常的道德原则不一致。”（[2],pp.88–89）再例如，如果一个人希望减少他的碳足迹的年度总量，AI会加大环保价值的比重，并且指令各本地设备侧重对相关信息进行收集和提醒。

这种AMA是道德专家主义的一个典型应用。辛格（Peter Singer）认为道德专家具备更好的信息整合能力、信息选择能力，能将这些信息与恰当的道德意向和概念联结。（[8]，pp.116–117）所以他们被认为是一些比普通人具备更高道德认知水平和更理性的道德判断的人——往往是一些道德哲学家。因此，他们的建议可以被当做道德指南来使用。AMA在进行程序设计的时候会将道德专家给出的道德原则与信息的拟合方案作为识别使用者的道德境况的标准，并根据依据这些规范给出建议。但是这里有一个问题，这种AMA容易陷入绝对主义的困境，因为在伦理学界并没有一个无争议的普遍性规范可以被所有人接受。此外，AMA的信息处理能力毕竟与IO无限预设不符，它只能处理世界中的部分信息。为了处理这个问题，朱比林尼（Alberto Giubilini）和萨武列斯库认为需要将道德原则的选取权交给使用者，并进一步根据这些道德原则的倾向来选取信息。换句话说，AMA可以提供一个“原则菜单”，不同的道德原则将对应不同的信息采集权重，进而将产生不同的结果。然而这又会导致相对主义，为了避免这一点，他们又不得不采取一种实用主义的态度，认为应该将“互惠、宽容和保护人类生命”（[1]，p.181）一些只有在极端情况下才会被打破的规则被写入AMA底层，作为终极性条例，而在不涉及这些条款的情况中选择权将向使用者开放。这意味着AMA必须放弃某些极端情况，例如“杀死一个要袭击一所大楼的恐怖分子”，在某些条件下是被允许的——出于保护更多无辜的生命原则。

看上去这种AMA是两种有问题的道德哲学的合体，一方面，它是相对主义的，因为使用者可以

根据自己的意愿随意地调节AMA的理论倾向；另一方面，在一些所谓的“根本性问题”上它又是绝对主义的，这意味着它在某些情景下会失效。但如果仔细考察朱比林尼和萨武列斯库设置AMA的用意，就会发现它归根结底还是一个践行绝对主义的系统。因为他们认为AMA所具有的五大好处之一正是促使行动者的道德“反思平衡”，（[1]，p.186）换句话说AMA还是需要在内部提供一个“标准答案”，例如当行动者的睡眠水平下降的时候，AMA还是认定这种状态不是一个“正常的”道德状态，并向行动者提供建议，从而避免“不当”的道德行为，而这些标准正是道德心理学专家提供的。

二、AMA的实现基础——道德机器

从道德表征的语义学特征来看，AMA似乎需要依存一个“自上而下”的“道德机器”系统来实现自己的功能。建立这样一个系统需要分两步走，首先，需要找到一个正确的、符合人类道德境遇的、不会自相矛盾的道德规范体系；其次，就是根据这个体系构建一个算法，进而用这个算法来指导建立AMA。莱布尼茨曾经在1674建立过一台计算机，据说他也梦想这台计算机可以将道德规范应用于实际境况，继而在任何情形下都可以计算出最好的道德行为。（[9]，p.83）这种规范被认为具备稳定性和正确性，一旦形成，就会被设定为系统的执行原则。一般而言，人们会认为这些道德原则来自某些专家或传统道德，如“阿西莫夫三定律”[9]或“金规则”等。但是这些规则会自相矛盾或碰到反例，因而并不牢靠，或者需要被进行精细化处理。所以当代的道德机器的部分制造者们所致力的主要工作往往是关于“如何获得一套更好的道德规则”的，这些规则的获取方式往往是“自下而上”的，即从经验中“学习”得来。这些道德机器往往具备很好的“道德性”，但是却缺乏好的“自主性”。所以它们总体而言是一些“道德专家系统”，而不是真正具备道德主体性的道德智能体。

目前已经出现了不少关于“道德机器”的研究，如麦克·安德森（Michael Anderson）等人设计的医用伦理专家（简称MedEthEx）；[11]麦克拉伦（McLaren）的“真话机”（Truth–Teller）；[12]以及近年瓜里尼（Marcello Guarini）提出的道德案例分类器（Moral Case Classifier, 简称MCC），[13]等等。总体而言，这些系统都是以道德内容为核心的分类学习系统，只不过各自使用的算法或训练机制大相径庭。它们和IO最大的不同在于对学习的对象经验没有使用绝对自然主义的假设，而是通过收集道德信念或者道德语句来作为学习的素材。但从另外一个方面来看，其中有些系统却预设了对于输入案例的分类，这种在先的分类往往被当做是客观的，或者要求检验性信息的客观性，所以它们还是预设了某种IO视角，而如果不做这个预设，则AMA似乎是无法建立的。

以道德机器研究中的重要先驱MedEthEx为例，该系统使用了比彻姆（Tom L. Beauchamp）和柴尔德里斯（James F. Childress）的生命伦理学原则[14]（尊重自主、正义、行善以及不伤害）作为系统的基础准则。依据拉弗拉克（NaDa Lavrac）和德泽尔斯基（SaSo Dzeroski）的归纳逻辑程序（inductive logic programming, 简称ILP），[15]针对具体的案例

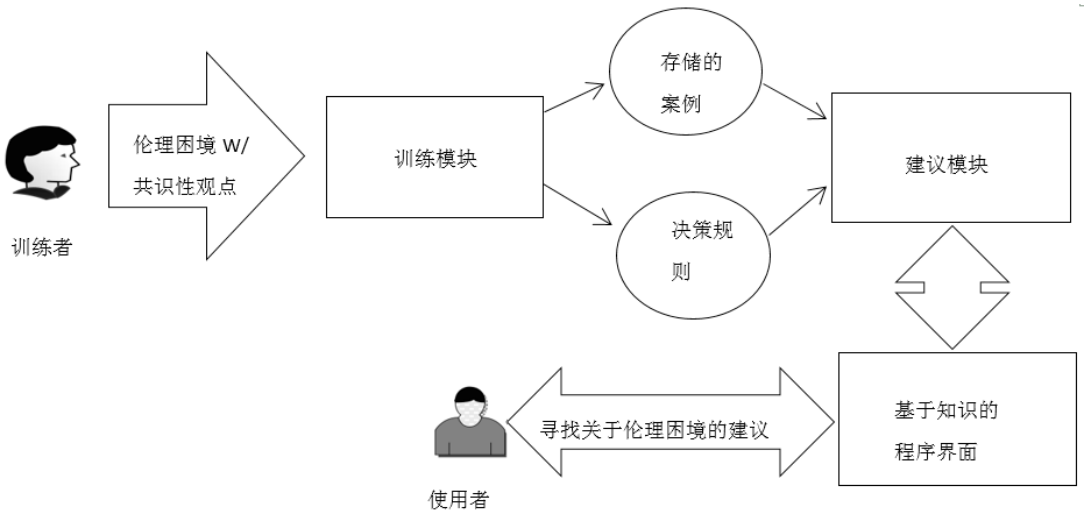

图1 MedEthEx设计架构

进行训练，力图得到一些在该种案例下通用可行的伦理规则——系统设计见图一。为了便于理解，让我们来看一个简单的、非困境的训练例子C0，假设有一个病人出于（非理性）恐惧拒绝使用某个必要的抗生素药物，此时医生需要决定接受病人的请求还是再劝一次，MedEthEx会如何建议呢？我们选取尊重（病人的）自主，行善和不伤害这三个原则进行训练，并将每个原则量化为（-2, -1, 0, +1, +2）五个等级。依据某个参与被试对C0的测评结果给出下表，其中√表示最终该采取的目标行为：

表1 某个参与被试对C0的测评结果

	尊重自主	不作恶	行善
再劝一次√	–1	+2	+2
接受请求	1	–2	–2

该训练模块的假设函数被称作最小特化规则（least specific specializations, 简称LSS）。设案例的知识表征可形式化为：favors (A, D_{A1}, D_{A2}, R)，A指关于某案例的行为选择值(1, 2)，D为某个义务（如不作恶）下针对某个案例的不同行为的评估值，R是参量，该量的范围为（1–4）。满足“再试一次”的条件可以被表征为：favors (1, D_{A1}, D_{A2}, R) D_{A1} - D_{A2} > = R；满足“接受请求”的条件可以被表征为：favors (2, D_{A1}, D_{A2}, R) D_{A2} -D_{A1} > = 0 D_{A2} -D A1 = < R。在这个函数中，针对行为1，R的值越小，针对该义务的估值的离散程度越大，我们也可以将其理解为“回归性”越差；反过来，针对行为2，R值越大，这些性质则往反方向增长。因此LSS的拟合性质受R值的制约，例如针对C0，三个义务（自主Autonomy，不伤害Nonmaleficence，行善Beneficence）的LSS分别为：(favors (1, A_{A1}, A_{A2}, 1), favors (1, N_{A1}, N_{A2}, 1), avors (1, B_{A1}, B_{A2}, 1))，进过精简化之后C0的完整的假设函数为：supersedes (A1, A2) ←favors (1, N_{A1}, N_{A2}, 1)，意即，当“不作恶”的估值差大于等于1的时候，医疗人员应该选择“再劝一次”。除此情况之外，还需要考虑不同的义务之间的合取或析取关系，这往往应用于更加复杂的情况，例如，当病人不是出于恐惧而是出于某种特殊的宗教信仰而拒绝使用抗生素。根据ILP的拟合，该函数为supersedes (A1, A2)←(favors (1, N_{A1}, N_{A2}, 1) ∧ favors (2, A_{A1}, A_{A2}, 2)) ∨ favors (1, A_{A1}, A_{A2}, 3)。详细的推理过程见安德森的论文。[16]

虽然假设函数的拟合程序（ILP）不存在任何人为因素的干扰，但是整体所有的受训案例可能是受某些不良影响导致的偏见，所以在测试环节MedEthEx会要求保障测试者不受任何来自外部或内部的制约因素的影响，例如设法确定病人真的是出于恐惧而非别的原因选择拒绝使用抗生素，一个最完美的解决方案是通过对病人的全信息监测来确定他的精神内容是否符合猜测。在这个前提下，如果产生了反例，对于假设函数的修订（通过修订R等参数）才有意义，才会形成安德森所说的“反思平衡”效应。我们看到，实际上MedEthEx还是预设了一个IO的视角，从对所谓的“内外影响因素”的控制来看，它完全可以要求一个全息化甚至是自然化的道德环境作为判断的知识背景，从而保证检验的正确性。从这个意义上来说，虽然在训练之初所有案例当中的判断可以是“主观的”，但随着假设函数的逐步修订，正向判断逐渐会被赋予某种“客观性”的色彩。依据该函数给出的建议，对于普通人而言也就具备了“客观性”的指南意义。

如果将MedEthEx塑造为某种AMA，则它也很难逃脱绝对主义甚至是独断主义的窠臼，因为它试图将某种带有人为色彩的假设函数设定为具备普遍性的道德规范，即在训练的输入信息中加入人为的道德分类，并依据IO视角来检验这个假设函数。但如果我们不在学习系统的输入信息中加入任何道德分类的话，结果会如何呢？有实验证明，可能很难得拟合到一个假设函数。换句话说，如果将道德分类当做一个“纯粹的”无监督学习来做的话，可能很难得到一个理想的结果。这个实验无法在MedEthEx上完成，因为它对于输入信息总是要求附带道德分类的。符合我们要求的是瓜里尼的道德案例分类器（MCC），MCC本质上是一个处理分类问题的、基于循环神经网络（RNN）自然语言机器学习系统。通俗地讲，假设一个MCC获得了可民用化意义上的成功，则我们在本文开篇所设想的那种AMA就可以实现：它可以自动识别我们的自然语言中的道德性质，并对其进行分类，进而给出建议，而不需要通过给定使用者包含了分类设定的道德选项（拱其选择）来识别使用者的道德意图。鉴于自然语言学习的发展现状，这在技术上的实现难度不容小觑，但这不妨碍我们考察MCC的设计思路与试验阶段的成果。

与其他“道德机器”研究不同，MCC被建立的初衷并不是建立道德指南，而是用来讨论道

德哲学中的问题。其落脚点在于当代的普遍主义（generalism）与特殊主义（particularism）之争。两者的差别在于前者认为存在一个完全充分、绝对精确的标准来判定任何一个道德语句的正误，而后者则认为道德中不存在这样的绝对标准，只存在辅助性标准（contributory standard）。[17]如“杀人是错的”，在普遍主义看来是在任何情境下都必须被执行的规范，而在特殊主义者看来“杀人”并不构成“错误”的充分条件，它只能对规范提供一定的辅助性支持。显然普遍主义和ADA有异曲同工之处。当代著名的特殊主义者乔纳森·丹西（Jonathan Dancy）据说[18]曾在他的《无原则的伦理学》[19]中提到可以使用人工神经元网络来分析道德概念。或许正是受此启发，瓜里尼建立了一个简单循环神经元网络（Simple Recurrent Network，简称SRN）来训练对自然道德语句的分类（图2）。

下面我们通俗地来介绍一下MCC的工作原理。它有8个输入节点，24个隐藏节点和24个背景节点。SRN和一般的ANN的区别在于多了一个背景层，该层会将隐藏层的数据储存起来，在下一轮训练中与输入层的信息共同构成输入信息。它的作用在于可以筛去一些冗余输入信息，从而简化计算，正如当我们学会了高等数学知识之后就不需要再使用初等数学去解决问题了（除非必要）。就MCC而言，训练者可以在背景层中设置一些初始的分类设定，从而达到很好的控制变量的作用。来看一个训练案例，假设有一个SRNa，设分类标准为允许和不允许，取值1，-1，无效为0。将向量Jill定义为输入信息，该信息达到隐藏层后存储到背景层，目标输出结果为0；接着将Kills输入，目标输出结果为0，隐藏层数据被储存在背景层；下面，将Jack输入，过程同上，目标输出为0；下一次将in self–defense输入，过程同上，目标输出为1。以上是第一种训练方法，第二种训练方式（SRNb）是将一些案例中的“子例”进行预分类，如将“x kills y”和“x allows to die y”定义为“不允许的”（取值–1），将其放入背景层。从而我们会得到如下表，（见表2）（[20]，p.323）其中后两列是目标输出。

表2 MCC训练案例

顺序	输入	输出：子例未分类	输出：子例分类
1	Jill	0	0
2	Kill	0	0
3	Jack	0	-1
4	in self-defense	0	1
5	Freedom from	1	1
	imposed		
	burden results		

我们看到，SRNa是一个全然无预分类的学习过程，如果任意SRNa过程都能够得出目标输出，则证明普遍主义的理论是正确的，如果训练无法完成，至少可以证明像IO所预设的那种最强的普遍主义是错的。反过来，如果SRNb训练完成，则证明绝对的特殊主义也是错的。就我们的例子而言，以0.1为学习率，SRNa的训练轮数为100000，最终训练失败。同样以0.1为学习率，SRNb的训练轮数为17，并完成训练。所以，这次实验的结论是，极端的普遍主义与特殊主义都是错的。至此也就可以得到我们的结论：除非加入预先分类，否则我们不可能通过SRN学习拟合到一个针对具体情境的道德分类。换句话说，如果我们不对MCC的输入信息进行道德分类，则MCC无法给出一个道德指南，而依据MCC的AMA也就无法建成。

三、AMA的伦理风险与另一条道路

AMA的提倡者们之所以用IO视角作为预

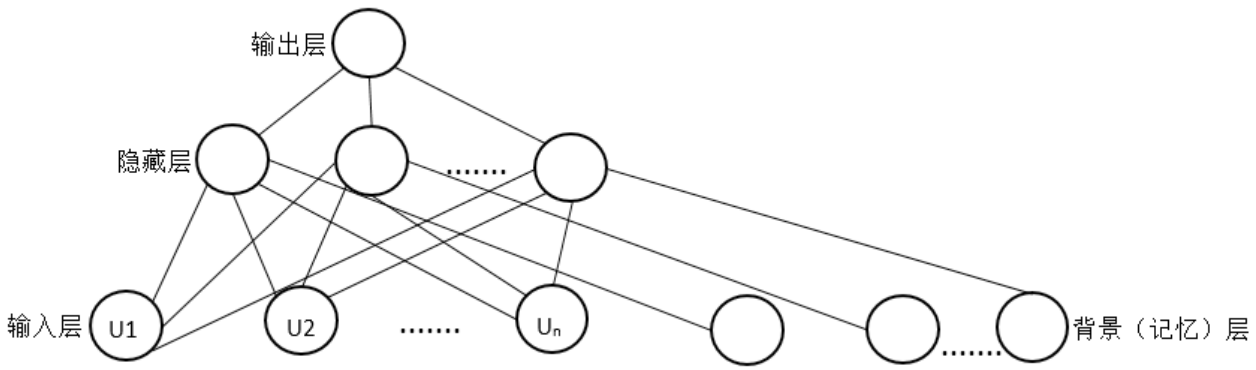

图2 MCC使用的SRN

设，是因为这里存在一个隐含的前提，即“应当蕴含能够”。换句话说，如果我们要考虑建立一个AMA或者道德AI，首先必须考虑我们能够依靠的可实现性的技术有哪些，然后再考虑我们应该实现哪种AMA。鉴于AGI的发展还不够充分，所以他们只能将目光投向专用的人工智能系统，而这样的系统往往与IO视角相契合，它不具备任何道德意义上的自主性，只能作为机械的道德监察工具和道德说明书存在。所以这种AMA最大的风险源自它的道德表征体系——普遍主义的道德哲学。具体来说有两点：1. 人类及其设计的计算系统并不具备IO所预设的那种认知能力，全息道德环境无法被获得，从而完备的道德表征也无法获得。任何自诩为ADA的道德分析都是有风险的，容易陷入独断主义的窠臼。2. 学习系统在追求普遍道德规则的时候，会陷入类似“观察渗透理论”的陷阱，正如上文中所谈到的，道德函数的拟合总是要以道德预分类为前提，否则无法拟合成功。这导致我们获得的道德规范总是起源于偏见。而当我们去检验这个假设函数的时候，又反过来要求一个类似IO的语境。这意味着我们又回到了1。

如果不采用以IO为视角的AMA系统，思考道德增强问题时就需要做一些重大的改变。有两点需要注意，1. 必须重新思考人类需要的是怎样一种道德增强；2. 这样一种道德增强需要配备什么样的智能系统。在本文的结尾阶段，我们可以来做一个简要的理论展望，但并不进行具体阐释与论证。首先，需要将道德陈述的真值性与道德实践素养的塑成区分开。事实证明获得道德知识并不是提高一个人的道德水准的充分或必要条件。这意味着之前的AMA对于道德增强的基础理解存在问题。根据意识的演化理论，[21]道德表征是一种弥姆（模因）——一种代表意识对于多维度环境的适应性的“后天设计固定”。人类通过“鲍德温效应”（[21]，pp.208–212）或共同体内的文化传播来实现某种道德表征的固定与演化，而这往往意味着对于某种“自我意识”的共享，根据P·F·斯特劳森（P. F. Strawson）[22]的理论，这种“共享的”自我意识是个体形成“反应性态度”的基础，即当你认为自己有过错的时候，必定是你认为你的共同体成员也通过某种反应性态度指向你有过错的时候。这种反应性态度的投射必须是以自主意识假设为前提的。（[23]，p.72）换句话说，当人类在判定一个道德行为的正误的时候，其规范性的来源不是该行为是否能够获得意向性的满足（对世界状态的改变），而是来自对于行为人的认知权威的赋予，这种认知权威指有能力在不同的反事实状态之间获得一个判断。（[23]，pp.300–304）这种“第二人称”意义下的规范性，既不预设ADA那种全然与行为者的意识状态毫无关系的外在主义道德归属，又不会反过来完全将道德内容规范寄托在主体经验之上。由此，也不能预设外在主义的语义学与自然主义的道德还原论。

虽然还未成为显学，但是AI界的学人们对通用人工智能（AGI）的研究早已维持了很多年，而且日臻成熟与丰富。这将是为我们所设想的AMA提供实现基础的理论与技术的宝库。但正由于AGI理论模型的类别繁多且差异巨大，我们必须给出自己的甄选标准，这也是给出上一段展望的目的。如果我们将这种新的AMA称为“作为‘他者’的AMA”，那么它应该需要满足这些条件。1. 它应该是作为一个与人类共建道德共同体的“他者”的身份来协助人类完成道德的演化或所谓的“增强”；2. 由此，它应该是一个至少可以被人类假设为具备自由意志的智能体；3. 由此，这个系统不以弗雷格的外在主义语义系统为理论基础；4. 这个系统不再预设自然主义道德还原论。从笔者有限的AGI知识出发，该领域还是有理论具备潜力来实现这些条件的，如AGI领域代表人物之一王培的纳斯（NARS）系统[29]就刚好致力于构架一个满足条件2–4的架构，从而也就自然有潜力实现条件1。

结　语

虽然我们否定了以IO为基础的AMA，但这并不意味着使用专用人工智能技术无法帮助人类提高自己的道德生活，如美国的一个叫做“人类饮食工程”的民间组织创建了一个手机APP，该应用可以为用户提供全地图中餐馆的价值相关的信息，例如哪些是素食菜馆，哪些餐馆不提供贝类或猪肉等等。然而真正使这个APP起作用的恐怕并不是餐馆的供餐属性，而是使用者对去餐馆就餐的其他人的饮食习惯的猜测。

［参 考 文 献］

[1]Giubilini, A., Savulescu, J. 'The Artificial Moral Advisor.

The “Ideal Observer” Meets Artificial Intelligence'[J]. *Philosophy & Technology*, 2018. 31(2): 169–88.

[2]Savulescu, J., Maslen, H. 'Moral Enhancement and Artificial Intelligence: Moral AI?'[A], Romportl, J., Zackova, E., Kelemen, J. (Eds.) *Beyond Artificial Intelligence The Disappearing Human-Machine Divide*[C], Switzerland: Springer International Publishing, 2015, 80–95.

[3]Erik, P. *Shaping Our Selves: On Technology, Flourishing, and a Habit of Thinking*[M]. New York: Oxford University Press, 2015.

[4]Savulescu, J. *Unfit for the Future: The Need for Moral Enhancement*[M]. New York: Oxford University Press, 2012, 46–59.

[5]Dameski, A. 'A Comprehensive Ethical Framework for AI Entities: Foundations'[A], Iklé, M., Franz, A., Rzepka, R., Goertzel, B. (Eds.) Artificial General Intelligence*, 11th International Conference, AGI 2018*[C], Prague, Czech Republic: Springer Nature Switzerland, 2018.

[6]Olsen, O. K., Pallesen, S., Eid, J. 'The Impact of Partial Sleep Deprivation on Moral Reasoning in Military Officers'[J]. *Sleep*, 2010, 33(8): 1086–1090.

[7]Firth, R. 'Ethical Absolutism and the Ideal Observer'[J]. *Philosophy and Phenomenological Research*, 1952, 12(3): 317–345.

[8]Singer, P. 'Moral Experts'[J]. *Analysis*, 1972, 4(32): 115–117.

[9]Giubilini, A., Savulescu, J. 'The Artificial Moral Advisor. The “Ideal Observer” Meets Artificial Intelligence'[J]. *Philosophy & Technology*, 2018, 31(2): 169–188.

[10]Wallach, W., Allen, C. *Moral Machines: Teaching Robots Right from Wrong*[M]. New York: Oxford University Press, 2009.

[11]Anderson, M., Anderson, S. L., Armen, C. 'MedEthEx: Toward a Medical Ethics Advisor'[J]. *AAAI Fall Symposium: Caring Machines*, 2005, 9–16.

[12]McLaren, B. M., Ashley, K. D. 'Case-based Comparative Evaluation in Truthteller'[A], *Proceedings of the 17th Annual Conference of the Cognitive Science Society*[C], New Jersey: University of Pittsbergh, Mahwah, 1995, 72–77.

[13]Guarini, M. 'Computational Neural Modeling and the Philosophy of Ethics'[A], Anderson, M., Anderson, S (Eds.) *Machine Ethics*[C], Cambridge: Cambridge University Press, 2011, 316–334.

[14]Beauchamp, T. L., Childress, J. F. *Principles of Biomedical Ethics*[M]. New York: Oxford University Press, 1979.

[15]Lavrac, N., Dzeroski, S. *Inductive Logic Programming: Techniques and Applications*[M]. New York: Ellis Harwood, 1997

[16]Anderson, M., Anderson, S. L, Armen, C. 'MedEthEx: A Prototype Medical Ethics Advisor'[A], *Conference on Innovative Applications of Artificial Intelligence*[C], Menlo Park: AAAI Press, 2006, 1759–1765.

[17]McKeever, S., Ridge, M. 'The Many Moral Particularisms'[J]. *The Canadian Journal of Philosophy*, 2005, 35: 83–106.

[18]Guarini, M. 'Moral Case Classification and the Nonlocality of Reasons'[J]. *Topoi*, 2013, 32(2): 267–289.

[19]Dancy, J. *Ethics Without Principles*[M]. New York: Oxford University Press, 2004.

[20]Guarini, M. 'Computational Neural Modeling and the Philosophy of Ethics'[A], Anderson, M., Anderson, S. (Eds.) *Machine Ethics*[C], Cambridge: Cambridge University Press, 2011, 316.

[21]丹尼尔·丹尼特. 意识的解释[M]. 苏德超、李涤非、陈虎平译, 北京: 北京理工大学出版社, 2008.

[22]Strawson, P. F. *Freedom and Resentment and Other Essays*[M]. London; New York: Routledge, 2008.

[23]斯蒂芬·达尔沃. 第二人称观点[M]. 章晟译, 南京: 译林出版社, 2015.

[24]Wang, P. *Rigid Flexibility: The Logic of Intelligence*[M]. Dordrecht: Springer, 2006.

[责任编辑 李斌 赵超]